Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 863592, 12 pages
http://dx.doi.org/10.1155/2013/863592

Research Article

Development of Conformation Independent Computational Models for the Early Recognition of Breast Cancer Resistance Protein Substrates

Melisa Edith Gantner,[1] Mauricio Emiliano Di Ianni,[1] María Esperanza Ruiz,[2] Alan Talevi,[1,3] and Luis E. Bruno-Blanch[1]

[1] *Medicinal Chemistry, Department of Biological Sciences, Faculty of Exact Sciences, National University of La Plata (UNLP), Argentinean National Council for Scientific and Technical Research (CONICET), CCT La Plata, Buenos Aires, B1900AJI La Plata, Argentina*
[2] *Quality Control of Medications, Department of Biological Sciences, Faculty of Exact Sciences, National University of La Plata (UNLP), Buenos Aires, B1900AJI La Plata, Argentina*
[3] *Biopharmacy, Department of Biological Sciences, Faculty of Exact Sciences, National University of La Plata (UNLP), 47 and 115, Buenos Aires, B1900AJI La Plata, Argentina*

Correspondence should be addressed to Alan Talevi; atalevi@biol.unlp.edu.ar

Received 30 April 2013; Accepted 25 June 2013

Academic Editor: Jielin Sun

ABC efflux transporters are polyspecific members of the ABC superfamily that, acting as drug and metabolite carriers, provide a biochemical barrier against drug penetration and contribute to detoxification. Their overexpression is linked to multidrug resistance issues in a diversity of diseases. Breast cancer resistance protein (BCRP) is the most expressed ABC efflux transporter throughout the intestine and the blood-brain barrier, limiting oral absorption and brain bioavailability of its substrates. Early recognition of BCRP substrates is thus essential to optimize oral drug absorption, design of novel therapeutics for central nervous system conditions, and overcome BCRP-mediated cross-resistance issues. We present the development of an ensemble of ligand-based machine learning algorithms for the early recognition of BCRP substrates, from a database of 262 substrates and nonsubstrates compiled from the literature. Such dataset was rationally partitioned into training and test sets by application of a 2-step clustering procedure. The models were developed through application of linear discriminant analysis to random subsamples of Dragon molecular descriptors. Simple data fusion and statistical comparison of partial areas under the curve of ROC curves were applied to obtain the best 2-model combination, which presented 82% and 74.5% of overall accuracy in the training and test set, respectively.

1. Introduction

ATP-binding cassette (ABC) efflux transporters comprise a diversity of active carriers which provide an efficient mechanism of defense against foreign chemicals (i.e., xenobiotics), among them drugs. To elicit the therapeutic response, drugs must often cross a number of cellular barriers, such as the gut wall and the capillaries endothelial cells. ABC efflux transporters limit drug absorption and distribution by translocating drugs from the cytoplasm to the cell exterior. These transporters are preferentially expressed at tissues that present barrier and/or excretory functions, for example, the intestinal wall, the canalicular membrane of hepatocytes in the liver, or the luminal membrane of the tubular cells in the kidney [1, 2], reducing the bioavailability of their substrates. Moreover, due to their wide substrate specificity, overexpression of such transporters is associated with cross-resistance phenomena to structurally unrelated drugs (multidrug resistance) in a wide range of diseases, from cancer to epilepsy [3–5]. Efflux transporters and metabolic enzymes seem to act in a coordinated or synergic manner, with the biotransformation products being often substrates for these

drug carriers [1]. Furthermore, metabolizing enzymes and efflux transporters are also upregulated in a coordinated manner by common nuclear receptors that, upon environmental chemical triggering agents, induce the expression of host defense systems towards potentially toxic chemical agents [1, 6, 7].

Even though P-glycoprotein (also known as ABCB1 or MDR1) was the first identified and is the most extensively studied member of the ABC superfamily, recent studies suggest that the effect of another member, breast cancer resistance protein (BCRP, or ABCG2) might have been underestimated in the past. A number of reports indicate that BCRP is the most abundantly expressed ABC efflux transporter in different segments of human intestine, both at mRNA [8, 9] and protein levels [10]. Similar observations have been found at the blood-brain barrier, where BCRP mRNA levels are around 8 times above those of P-glycoprotein and represent 85% of the total ABC transporters mRNA [11], while at the protein level, BCRP levels are about 1.6 times higher [12, 13]. Therefore, regulation of BCRP and/or early recognition of BCRP substrates are critical aspects to optimize oral drug absorption, increase drug bioavailability, and design novel therapeutics aimed at brain conditions and diseases linked to BCRP-mediated multidrug resistance issues (e.g., cancer).

The most advanced research regarding ABC transporters modulation relates to add-on therapies of specific inhibitors of ABC transporters, a strategy that was originally conceived for cancer treatment. Although preclinical and initial clinical results with first- and second-generation inhibitors have been encouraging, some trials stopped at phase III due to serious adverse effects [3, 5, 14–16]; such outcome has put in doubt the strategy of overcoming cellular drug resistance by the use of transporters inhibitors, even though trials continue in order to find more effective and safe inhibitors for P-glycoprotein and other drug carriers [16]. It is worth highlighting that ABC transporters comprise a concerted, complex transport system whose substrates are not only drugs, but also endogenous compounds (e.g., waste products, bile salts) and toxins. Thus, their permanent impairment or disruption is likely to result in severe side effects (especially in those therapeutic backgrounds that demand long-term treatment). Recent research has then focused on elucidating intracellular signaling pathways that control ABC transporters (their expression, intracellular trafficking, activation, and inactivation). It is proposed that finding the molecular switches of these transporters will allow selective modulation of transporters function and/or expression for therapeutic purposes in different clinical scenarios [17], which includes turning off the efflux mechanisms for short, controlled periods of time. Other alternatives, probably safer approaches propose avoidance of substrate-transporter interaction by the encapsulation of therapeutic agents within nanosystems (a "Trojan horse" approach) [18], or designing drugs or prodrugs which are not recognized by drug carriers [16, 19, 20].

At present, limited studies have been done to develop high throughput *in silico* models for the early identification of BCRP substrates, in order to assist virtual screening and computer-aided design of novel BCRP nonsubstrates therapeutics. Recently, Hazai et al. obtained a support vector machine model based on 5 Dragon descriptors [21]. To that purpose, they compiled a 263-compound wild-type BCRP substrates and nonsubstrates dataset which was randomly partitioned into a 167-compound unbalanced training set (it contained far more substrates than nonsubstrates), a 56-compound test set, and a 40-compound independent external set. The model showed an overall accuracy of 76% on the training set, 75% on the test set, and 72.5% on the external set; moreover, it presented much more accuracy in the identification of substrates than nonsubstrates (a possible consequence of the unbalanced training set). Some of the descriptors incorporated into this model were 3D (conformation dependent) structural features, which implies that considerable preprocessing (conformational analysis) of the predicted structures is needed before proceeding to prediction itself, which may hamper the screening efficiency of the algorithm, especially if we take into account that available public databases for virtual screening purposes (e.g., Drugbank, ZINC database) compile thousands to millions of chemical compounds. This same issue can be envisaged for the 17-descriptor model reported by Zhong et al. [22], who combined genetic algorithms and support vector machines to obtain, from a more limited unbalanced 177-compound dataset of BCRP substrates and nonsubstrates (again, randomly partitioned into training and test sets), a model with 85% overall accuracy. Once more, this model is majorly composed of 3D descriptors.

It has been pointed out that the polyspecificity/broad substrate specificity of ABC transporters due to multiple separate binding sites or "binding zones," binding sites accommodating more than one ligand and high protein flexibility determine a complex phenomenon which can only be partially addressed by current methods in the computational drug design field [23, 24]. This explains why many modeling efforts to identify ABC transporters substrates have resorted to ensemble learning or locally weighted methods [25–28]. In fact, BCRP presents at least two binding sites [29–31]. Here, we present the development of an ensemble of linear classificatory models capable of differentiating BCRP substrates and nonsubstrates.

Contrasting the previously discussed models, our ensemble is entirely based on conformation independent descriptors, which makes it an adequate *in silico* filter to assist virtual screening campaigns in a highly efficient manner. The models have been derived from a relatively large 262-compound dataset which was rationally partitioned—through combined hierarchical and k-means clustering—into a representative and balanced 164-compound training set (85 substrates and 79 nonsubstrates) and a 98-compound test set (71 substrates and 27 nonsubstrates). Furthermore, on the basis of receiving operating characteristic (ROC) curves analysis, the score threshold can be optimized to prioritize the accuracy in the prediction of either substrates or nonsubstrates, depending on background-dependent criteria. In order to minimize the early selection of BCRP substrates as drug candidates, we have prioritized substrate accuracy. Such decision was supported by statistical comparison of the partial area under the curve (AUC) of ROC curves.

2. Materials and Methods

2.1. Dataset. A 305-compound diverse dataset containing BCRP substrates and nonsubstrates was compiled from the literature. It is known that a single-nucleotide substitution at R482 can modify the affinity of BCRP for substrates [32–39]; however, the clinical consequences of such variant are not clear to the moment [40]. Therefore, from the original 305-compound dataset, we selected 262 compounds which are substrates and nonsubstrates of human wild-type BCRP, the subject of this modeling effort. BCRP substrates or nonsubstrates of BCRP homologs from other species with no evidence of human BCRP-mediated transport were not included to avoid noise due to inter-species variability in substrate specificity. The dataset was split into a 164-compound balanced training set (85 substrates, 79 nonsubstrates) and a 98-compound independent test set reserved for external validation. In order to obtain representative partitions of the dataset compounds, a combined hierarchical and k-means clustering approach was applied. The LibraryMCS v0.7 (ChemAxon) hierarchical clustering approach was applied in combination with the k-means clustering as implemented in Statistica 10 cluster analysis module (Statsoft Inc., 2011). LibraryMCS relies on the maximum common substructure (MCS, i.e., the largest subgraph shared by two chemical graphs) to cluster a set of chemical structures. The algorithm applies similarity search to the pool of molecules, and the two structures with the highest similarity coefficient are considered more likely to share a large MCS. Once this probable MCS has been established, substructure search is carried out in order to find the MCS of multiple structures efficiently, without exhaustive pairwise comparison. Certainly, it is possible that the two structures with highest similarity coefficient are not the ones that share the largest MCS; thus, library MCS leads to reproducible but approximate solutions [41]. As suggested by Everitt et al. [42], hierarchical clustering has been applied here to define an initial partition of n objects into g groups, selecting the smallest common substructure of 9 atoms. The groups of compounds were later optimized by k-means algorithm, minimizing the Euclidean distance to the group centers. A series of descriptors computed with Dragon 4.0 (Milano Chemometrics, 2003) representing different aspects of molecular structure (namely, molecular weight, log P, polar surface area, number of H bonds acceptors and donors, total information index of atomic composition, sum of atomic van der Waals volumes, sum of atomic Sanderson electronegativities, and 2D Petitjean shape index) were normalized and applied to calculate such distance. Once the clusters were separately identified in the substrates and nonsubstrates classes, around 50% of each cluster from the substrates category and 25% of each cluster in the nonsubstrates category were assigned to an independent test set for validation purposes, while the remaining percentage of the clusters was retained as training set for modeling purposes. This scheme allowed obtaining a balanced training set where, unlike in previous modeling efforts, neither the substrates nor the nonsubstrates were markedly overrepresented. The structures of both training and test set compounds are provided as pdf files in Supplementary information available online at http://dx.doi.org/10.1155/2013/863592 so that the reader can appreciate the structural diversity of the dataset. Representative members of each cluster in the substrate and non-substrate categories are shown in Figure 1.

2.2. Molecular Descriptor Calculation and Modeling Method. Dragon software for molecular descriptors calculation, version 4.0 (Milano Chemometrics, 2003) was used for the calculation of 867 low-dimensional (0D–2D) descriptors, distributed along 12 blocks of descriptors, for example, constitutional descriptors, topological descriptors, connectivity indices, Galvez topological charge indices, functional groups count, and others. Since such a high number of descriptors may result in chance correlations between the modeled property and a subset of descriptors, 102 subsets of descriptors obtained from random combinations of the blocks of Dragon low-dimensional descriptors were considered as independent pools of descriptors, each combination containing around 200 molecular descriptors. For example, the first pool of descriptors (180 descriptors) emerged from combination of the following Dragon blocks of descriptors: constitutional descriptors, eigenvalue-based indices, 2D autocorrelations and molecular properties; the second pool of descriptors (195 descriptors) combined walk and path counts, connectivity indices, functional group counts and BCUT descriptors, and so on. The use of this strategy, called *random subspace*, has proved to be effective for ensemble learning to combine weak learners in order to obtain strong learners [43, 44]. Descriptors with constant or near-constant values for the training set associated to low information content were removed from descriptors pools.

A binary, dummy variable codifying the category of each compound was used as dependent variable (class = 1 for substrates and class = −1 for nonsubstrates). Stepwise forward multiple linear regression was used to select the descriptors from each random pool that best discriminated the category of the compounds. Obtaining all possible descriptors subsets would demand $D!/[d!(D - d)!]$, where D is the number of descriptors in a given descriptor pool and d is the number of descriptors included in a given model. This is very computationally demanding or even unfeasible when D is large, as in the present work. Therefore, we resort to a stepwise approach which, although faster, leads to suboptimal solutions.

Linear discriminant analysis (LDA) was used to characterize the correspondent linear discriminant functions (dfs). Dfs assume the following general form:

$$\text{df value} = a_0 + \sum_i a_i - d_i, \tag{1}$$

where a_0 is constant and a_i is the coefficient associated with molecular descriptor d_i. Due to the values arbitrarily assigned to substrates and nonsubstrates, substrates will tend to have positive df values, and nonsubstrates will tend to assume negative values.

The binary classification scheme reduces the error associated with combining data obtained in different labs and conditions [45]. Multiple regression and discriminant analysis

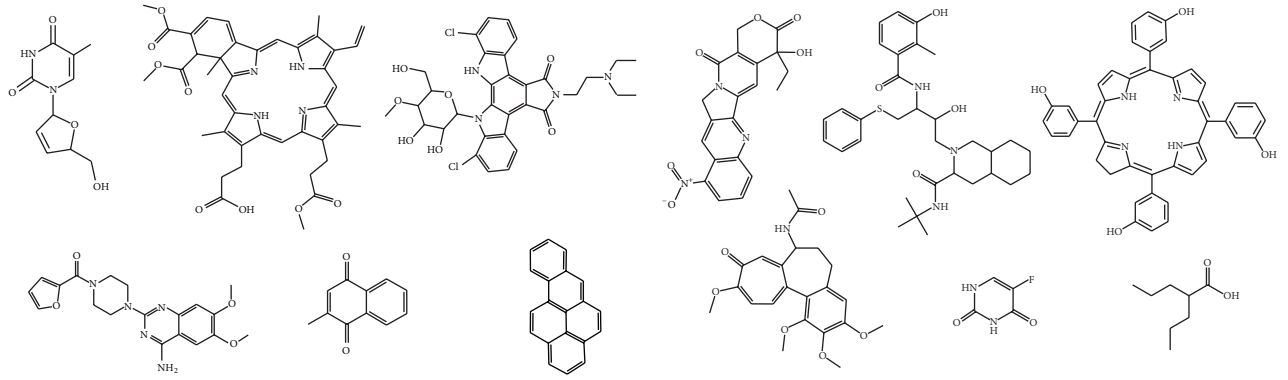

Figure 1: Representative BCRP substrates (left) and nonsubstrates (right) from the six most populated clusters in the dataset.

modules from Statistica 10 were used for modeling purposes. Tolerance values no lower than 0.1 were used in order to avoid inclusion of highly correlated pairs of descriptors. The minimum cases to predictors ratio allowed was 11 (11 or more cases in the training set for each descriptor included in the model) in order to reduce chances of overfitting; thus, models including at most between 10 and 15 descriptors were obtained through a stepwise forwards procedure. Only descriptors with significant coefficients at an alpha level of 0.05 are allowed into the model. Randomization, stratified leave-group-out (LGO) cross-validation and external validation (predicting the class for the independent 98-compound test set) were used to assess all models robustness and predictive ability. 50 randomized models were built in the randomization test. In each LGO row, 10 compounds were randomly removed from the training set, and the resulting LGO models were used to assess the category of the removed compounds; this process was repeated 50 times, checking that all the compounds in the training set had been removed in at least one LGO round.

2.3. Combining Models. Two important indicators of the performance of a given QSAR model are sensitivity (Se) and specificity (Sp). They are defined by the following expressions:

$$\begin{aligned} \mathrm{Se} &= \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \\ \mathrm{Sp} &= \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}, \end{aligned} \tag{2}$$

where TP refers to true positives, FN refers to false negatives, TN refers to true negatives, and FP refers to false positives. Here, since we are looking for compounds that are not transported by BCRP (BCRP nonsubstrates) and we want to discard BCRP substrates, the previous expressions may be rewritten as follows:

$$\begin{aligned} \mathrm{Se} &= \frac{\text{True nonsubstrates}}{\text{True nonsubstrates} + \text{False substrates}}, \\ \mathrm{Sp} &= \frac{\text{True substrates}}{\text{True substrates} + \text{False nonsubstrates}}. \end{aligned} \tag{3}$$

By modifying the selection threshold from the lowest to the highest score provided by the individual models or the model ensemble, Se and Sp will evolve in opposite ways. Consequently, it is not possible to optimize both parameters simultaneously, and a tradeoff has to be found. ROC curves are a wide-used tool to assess and compare the performance of different models [46]. They are graphical plots of the sensitivity (true positives rate) versus 1 minus specificity (i.e., 1 less the false positives rate), for a binary classifier system, as its discrimination cutoff value changes. ROC curves provide a rational and user-friendly basis to balance type I and type II errors, selecting optimal models and optimal cutoff values. The AUC can be used for general comparison purposes of different models or methodologies. An ideal model will present an area under the ROC curve of 1 (equivalent to perfect classification, i.e., a sensitivity of 1 and a specificity of 1 for a given cutoff value), while random classification is represented by a line of slope 1 and corresponds to an area under the ROC curve of 0.5. Here, we have built ROC curves to compare the performance of the individual models developed and the performance of 2-model ensembles obtained through simple data fusion schemes.

It has been pointed out that there is no general rule for balancing errors [46]. Balancing FP and FN depends on pragmatic considerations that are to be judged by the researcher [47]. We are interested in adopting a conservative attitude and developing highly specific models, that is, models capable of discarding practically all BCRP substrates. This is strongly related to our background: a small academic research group from a developing country with limited resources to invest in drugs acquisition and pharmacologic testing. Therefore, we will privilege Sp over Se. At the risk of losing some valuable scaffolds when applying our models in virtual screening campaigns, we will choose to avoid acquiring or synthesizing a drug candidate that, once send to pharmacological testing, will prove to be a FP (a drug that was predicted as a BCRP nonsubstrate but which is actually transported by BCRP). Taking into account that BCRP is characterized by broad substrate specificity (probably because, in part, of the existence of multiple binding sites in the protein), we

have chosen to look for combinations of dfs that provide the lowest rate of FP in the external validation. Substrate polyspecificity indicates that it might be difficult to obtain a single model that is capable of identifying the entire set of BCRP substrates. We have combined the models by the very simple strategy of exhaustively looking for all the possible 2-model combinations of the models built from each of the pools of descriptors. The maximum (MAX operator) values among the values provided for each compound by the independent classifiers that compose the ensemble and the average (AVE) of the two values provided for each compound by the two independent models that compose the ensemble were used as data fusion schemes.

2.4. Statistical Analysis of ROC Curves. It has been suggested that AUC of the ROC curves may not be the best parameter to compare different models, especially if the modeler is interested in comparing a particular region of the ROC curve instead of the entire curve (in our case, e.g., we are interested in the early zones of the curves corresponding to high Sp). A key requirement for the success of virtual screening is the ability of the combination of a scoring function to rank actives early in a large set of compounds; this has been referred as the early recognition problem [48]. Therefore, to compare the performance of the individual models and the model ensembles, we have applied, along with the total AUC comparison, partial AUC (pAUC) statistical analysis [49] and the calculation of other metrics (enrichment descriptors) that have been proposed to address the early recognition problem in virtual screening [48, 50, 51].

To fit ROC curves to the three sets of data, a nonparametric approach [52, 53] was used. The analysis was performed using *roc.jar* [54] and *pROC* packages for R [55].

The empirical Se and Sp were derived by dichotomizing the observed (empirical) values into positive or negative test-results for each observed cut point of the variable (y). As y varies over the observed values of the variables, the empirical ROC curve is defined as the discrete set of Se(y) and [1 − Sp(y)] values joined by straight lines [56]. The curve passes through point (0, 0) when y is larger than the maximum value observed, and it monotonically increases to the point (1, 1), as y decreases to the smallest observed value. To be informative, the curve should be above the 45° line at least for some of the values, where Se(y) is equal to 1 − Sp(y) [57].

In the nonparametric approach, the AUC is estimated by the trapezoid defined by the empirical set of Se(y) and [1 − Sp(y)] values, and its value is related to the U statistic for the two-sample Mann-Whitney/Wilcoxon rank-sum test [52, 53] and can be estimated accordingly, as well as the confidence intervals (CI) and the variance-covariance matrix [58]. Since the three ROC curves were built on variables observed on the same sample, they were paired, and therefore the null hypotheses of equal AUC's between two of them (A and B in the formula) were tested with a two-sample z-test:

$$z = \frac{\left(\mathrm{AUC}_A - \mathrm{AUC}_B\right)}{\sqrt{\mathrm{Var}_A + \mathrm{Var}_B - 2 \cdot \mathrm{Covar}_{AB}}}. \quad (4)$$

For a pAUC estimation between two given Sp values, pROC package [56] was used.The numerical value of pAUC is estimated by the trapezoidal rule, whereas significance testing and confidence interval computation is performed by bootstrap with nonparametric stratified resampling (n = 2000). In stratified bootstrap, each replicate contains the same number of cases and controls than the original sample. Stratification is especially useful if one group has only little observations, or if groups are not balanced [59].

For the evaluation of the model performance, several enrichment descriptors were calculated for each model: area under the accumulation curve (AUCc) [60], enrichment factor (EF) [61], robust initial enhancement (RIE) [62] and Boltzmann-enhanced discrimination of ROC (BEDROC) [48]. All of these metrics were calculated with the *enrichvs* package for R [63].

2.5. Simulated Virtual Screening Campaign. An issue that emerges from using a reduced dataset (such as our 98-compound test set) is that the enrichment metrics derived exhibit a higher variance compared to significantly large datasets. Experiments conducted by Truchon and Bayly [48] show that the standard deviations associated with enrichment metrics such as ROC or AUCc are higher for small datasets and converge to a constant value when the size of the dataset increases.

The other problem is related to the high ratio of actives which mainly hinders the early recognition ability in what is known as the "saturation effect." That is, for datasets with a high ratio of hits (in our case, BCRP nonsubstrates), once hit compounds saturate the early part of the ordered list, the enrichment metric cannot get any higher. To estimate in a more realistic way the utility of our model in a real virtual screening approach, we have dispersed our test set among 479 putative substrates acting as decoys. Such putative substrates are substrates of BCRP from other species rather than human or highly similar compounds to human BCRP substrates which have been retrieved from either ZINC database or Pubchem. This simulated database thus contains 27 known nonsubstrates among 550 known or putative substrates among 550 known or putative substrates; that is, the nonsubstrates ratio is less than 0.05, representing a more challenging set to assess the enrichment ability of our models. Note that some of these putative substrates (decoys) might actually be nonsubstrates; thus, the true performance of our models may be even higher than the one obtained through this simulated experiment.

3. Results and Discussion

Based on the considerations exposed in the previous section, varying tolerance values between 0.1 and 0.5 and the maximum number of steps in the stepwise forward procedure between 10 and 15, we obtained 196-individual models from the 102-descriptor pools. Exhaustive 2-model combinations of these 196 models were performed by applying the MAX and AVE data fusion strategies. Table 1 presents the statistics and validation outcome for those individual models that were

Table 1: Features of the best individual model (Model 1) and the other individual models (models 2 to 4) that composed the two best 2-model ensembles.

Descriptors included	F	P value	Sp training set*	Se training set*	Overall accuracy training set*	Sp test set*	Se test set*	Overall accuracy test set*	Leave-group-out CV[1]	Randomization[2]
Model 1: *m*log *P*2 (squared Moriguchi octanol-water partition coefficient), nCrR2 (no. of ring quaternary C(sp3)), JGI7 (mean topological charge index of order 7), nCONHR (no. of secondary amides (aliphatic)), nHAcc (no. of acceptor atoms for H-bonds (N O F)), and GGI8 (topological charge index of order 8).	8.04	<0.000000	79%	68%	74%	63%	74%	66%	70.4% (±11.9)	64.4% (±3.4)
Model 2: BEHm2 (highest eigenvalue no. 2 of Burden matrix/weighted by atomic masses), BELe2 (lowest eigenvalue no. 2 of burden matrix/weighted by atomic Sanderson electronegativities), Hy (hydrophilic factor), LAI (Lipinski alert index), LP1 (Lovasz-Pelikan index), BEHp1 (highest eigenvalue no. 1 of Burden matrix/weighted by atomic polarizabilities), SEigp (eigenvalue sum from polarizability weighted distance matrix), and VRA2 (average Randic-type eigenvector-based index from adjacency matrix).	7.52	<0.000000	75.3%	74.7%	75%	76%	66.7%	73.5%	67% (±15)	61.5% (±3.6)

Table 1: Continued.

Descriptors included	F	P value	Sp training set*	Se training set*	Overall accuracy training set*	Sp test set*	Se test set*	Overall accuracy test set*	Leave-group-out CV[1]	Randomization[2]
Model 3: D/Dr11 (distance/detour ring index of order 11), nCONHR, nCO (no. of ketones (aliphatic)), X0Av (average valence connectivity index chi-0), nCaH (no. of unsubstituted aromatic C(sp2)), Xt (total structure connectivity index), PW4 (path/walk 4-Randic shape index), D/Dr12 (distance/detour ring index of order 12), T(O..O) (sum of topological distances between O..O), nNHRPh (no. of secondary amines (aromatic)), SPI (superpendentic index), and Rww (reciprocal hyperdetour index).	10.39	<0.000000	83.5%	83.5%	83.5%	73.2%	74%	73.5%	81.2% (±11.3)	62.4% (±5.1)
Model 4: $m\log P$2, JGI7, SRW10 (self-returning walk count of order 10), piPC02 (molecular multiple path count of order 02), and Hy.	6.56	<0.000014	63.3%	70.6%	67%	77.5%	70.4%	75.5%	64.8% (±13.6)	58.3% (±4.05)

*Considering zero as a cutoff value between substrates and non-substrates. This threshold may be later optimized through ROC curves analysis to provide a background-dependent optimal balance between Sp and Se.

[1]Results are presented as the average result for the folds ± the standard deviation.

[2]Results are presented as the average performance of the randomized models ± the standard deviation.

Table 2: Features of the best individual model (Model 1) and the best ensembles selected.

Model/ensemble	AUC ROC curve training set	AUC ROC curve test set	Sp training set*	Se training set*	Overall accuracy training set*	Sp test set*	Se test set*	Overall accuracy test set*
Model 1	0.796	0.748	78.8%	68.3%	74%	63.4%	74%	66%
Ensemble 1	0.850	0.785	83.5%	74.7%	79%	70.4%	74%	71.4%
Ensemble 2	0.902	0.804	84.7%	79.7%	82%	76%	70.4%	74.5%

*Considering zero as a cutoff value between substrates and non-substrates.

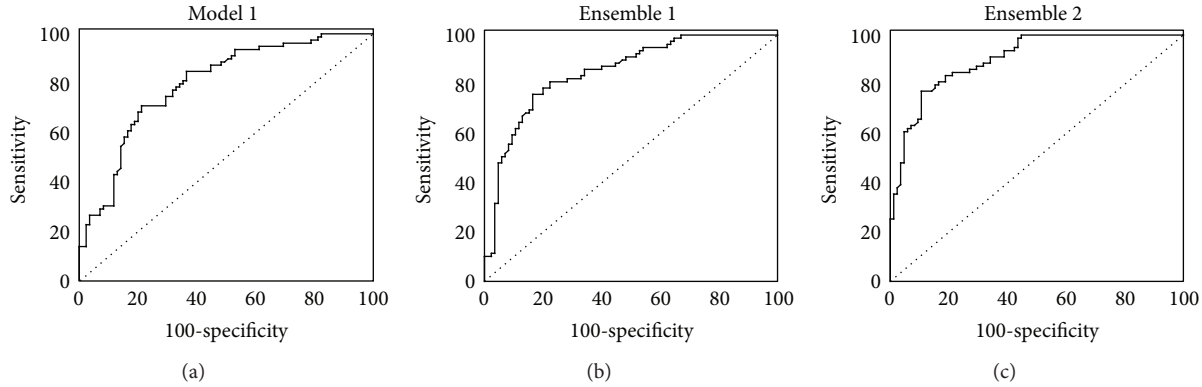

Figure 2: ROC curves of the training set for the best individual model plus the two best model ensembles.

later selected in the best 2-model ensembles. All the individual models that take part in the best 2-model combinations present an excellent cases per predictor ratio (from 13.7 to 32.8) indicating very low chance of overfitting. They present tolerances between 0.1 and 0.2, suggesting low pairwise correlation between the descriptors included in the models. As expected, the explanatory power of the randomized models is significantly below that of the actual (nonrandomized) model, since the correlation between the molecular structure and the modeled property is abolished when the dependent variable (in our case, the class label) is scrambled among the training set compounds. All models present conformation independent descriptors; thus, they may be applied to assist virtual screening campaigns, detecting (and discarding) potential BCRP substrates at the early stages of drug development projects, so that the retained candidates do not present low bioavailability or drug interactions issues due to their efflux transport by BCRP. The results of both the internal cross-validation and the external validation show adequate predictive power, especially considering the broad substrate specificity of BCRP. Nevertheless, these results have been remarkably improved when combining the individual models into 2-model ensembles, as shown in Table 2.

The AVE data fusion scheme outperformed the use of the MAX operator. This is in line with previous reports empirically showing that consensus prediction by simple averaging of outputs of individual models is an efficient way to enhance predictive performances [64–69].

The statistical comparison of the ROC curves (Figure 2) proved that ensemble 2 outperforms the best individual model (model 1) in the high Sp region in both the training set and the simulated 577-compound database. The results are summarized in Tables 3 and 4. We are particularly interested in the high Sp regions in order to assure discarding BCRP substrates in the early stages of the drug discovery process. The results confirm the utility of ensemble learning to identify ABC transporters substrates and nonsubstrates. ROC curves may also be applied to optimize the score threshold of the model or model ensemble, in order to select the best possible balance between Sp and Se, taking into account the active yield and other background-dependent considerations (e.g., budget, need to identify novel scaffolds).

4. Conclusion

We have developed a 2-model ensemble based on conformation independent Dragon descriptors for the identification of BCRP substrates and nonsubstrates. Since the descriptors incorporated into the models do not require preprocessing of the predicted chemical structure, the ensemble is particularly suitable to be applied in virtual screening campaigns of large chemical libraries in a highly efficient manner. Statistical comparison of ROC curves indicates that the best 2-ensemble model outperforms the best individual models generated. One should keep in mind that the broad substrate specificity of BCRP (and, in general, ABC transporters) makes it difficult to find a single linear relationship capable of accurately classifying substrates and nonsubstrates, a fact that justifies the application of more complex strategies such as ensemble learning or locally weighted regression methods. Unlike previously developed modeling efforts towards recognition

Table 3: Results of the calculation of the total and partial areas under ROC curve for the best individual model and the 2 best 2-model ensembles.

	Model 1	Ensemble 1	Ensemble 2
Training set			
Total ROC curve AUC (95% CI)	0.7964 (0.7284–0.8643)	0.8503 (0.7917–0.9089)	0.9022* (0.8578–0.9466)
Partial ROC curve AUC (±SD)			
From 1 to Sp = [1 to 0.70]	0.1612 (±0.0199)	0.1877 (±0.0199)	0.2218 (±0.0163)*
From 1 to Sp = [1 to 0.75]	0.1252 (±0.0171)	0.1459 (±0.0178)	0.1771 (±0.0145)*
From 1 to Sp = [1 to 0.80]	0.0917 (±0.0147)	0.1059 (±0.0149)	0.1337 (±0.0122)†
Simulated 577-compound database			
Total ROC curve AUC (95% CI)	0.7321 (0.6413–0.8229)	0.7357 (0.6418–0.8297)	0.7707 (0.6746–0.8669)
Partial ROC curve AUC (±SD)			
From 1 to Sp = [1 to 0.70]	0.1035 (±0.0213)	0.1127 (±0.0223)	0.1421 (±0.0223)
From 1 to Sp = [1 to 0.75]	0.0708 (±0.0183)	0.0794 (±0.0189)	0.1075 (±0.0208)†
From 1 to Sp = [1 to 0.80]	0.0458 (±0.0140)	0.0504 (±0.0148)	0.0765 (±0.0162)†

*The value is different from the best individual model (model 1) ($P < 0.001$).
†The value is different from the best individual model (model 1) ($P < 0.01$).

Table 4: Results of the enrichment parameters calculation for the best individual model and the best 2-model ensemble.

	Model 1	Ensemble 2
Training set		
Accumulation curve AUC (AUCc)‡	0.6458	0.6938
Enrichment factor (EF)	1.9294	1.9294
Robust initial enhancement (RIE)	1.8338	1.9261
Bedroc	0.9505	0.9983
Simulated 577-compound database		
Accumulation curve AUC (AUCc)‡	0.7212	0.7581
Enrichment factor (EF)	2.9630	5.9259
Robust initial enhancement (RIE)	2.9455	4.6663
Bedroc	0.2268	0.3593

‡It verifies that ROC AUC = AUCc/$R_i - R_a/(2 * R_i)$, where R_i and R_a are the ratios of inactives and actives, respectively.

of BCRP substrates and nonsubstrates, our models have been derived from a relatively large dataset which was split into representative training and test set through clustering analysis; the obtained training set presents a fair balance between the number of substrates and nonsubstrates. The studied ensemble is a potentially valuable tool to assist virtual screening and computer-aided drug design campaigns, as suggested by the outcome of the simulated virtual screening campaign. With the help of the constructed ROC curves, Sp and Se may be balanced to attend specific user requirements.

Conflict of Interests

The authors declare no conflict of interests related to the present paper.

Acknowledgments

M. E. Gantner, M. E. Di Ianni, and M. E. Ruiz are CONICET fellowship holders. A. Talevi is a Member of the Scientific Research Career at CONICET. L. E. Bruno-Blanch is a Researcher of Facultad de Ciencias Exactas, Universidad Nacional de La Plata. The authors would like to thank UNLP (Incentivos X-597), CONICET (PIP 11220090100603), and ANPCyT (PICTs 2010-2531 and 2010-1774) for providing funds to develop our research.

References

[1] L. M. S. Chan, S. Lowes, and B. H. Hirst, "The ABCs of drug transport in intestine and liver: efflux proteins limiting drug absorption and bioavailability," *European Journal of Pharmaceutical Sciences*, vol. 21, no. 1, pp. 25–51, 2004.

[2] C. G. Dietrich, A. Geier, and R. P. J. Oude Elferink, "ABC of oral bioavailability: transporters as gatekeepers in the gut," *Gut*, vol. 52, no. 12, pp. 1788–1795, 2003.

[3] H. Potschka, "Role of CNS efflux drug transporters in antiepileptic drug delivery: overcoming CNS efflux drug transport," *Advanced Drug Delivery Reviews*, vol. 64, pp. 943–952, 2012.

[4] Z.-S. Chen and A. K. Tiwari, "Multidrug resistance proteins (MRPs/ABCCs) in cancer chemotherapy and genetic diseases," *FEBS Journal*, vol. 278, no. 18, pp. 3226–3245, 2011.

[5] A. K. Tiwari, K. Sodani, C.-L. Dai, C. R. Ashby Jr., and Z.-S. Che, "Revisiting the ABCs of multidrug resistance in cancer

chemotherapy," *Current Pharmaceutical Biotechnology*, vol. 12, no. 4, pp. 570–594, 2011.

[6] B. Bauer, A. M. S. Hartz, J. R. Lucking, X. Yang, G. M. Pollack, and D. S. Miller, "Coordinated nuclear receptor regulation of the efflux transporter, Mrp2, and the phase-II metabolizing enzyme, GSTπ, at the blood-brain barrier," *Journal of Cerebral Blood Flow and Metabolism*, vol. 28, no. 6, pp. 1222–1234, 2008.

[7] O. Burk, "Nuclear receptor-mediated regulation of drug transporters," in *Nuclear Receptors in Drug Metabolism*, W. Zie, Ed., John Wiley & Sons, New York, NY, USA, 2009.

[8] G. Englund, F. Rorsman, A. Rönnblom et al., "Regional levels of drug transporters along the human intestinal tract: co-expression of ABC and SLC transporters and comparison with Caco-2 cells," *European Journal of Pharmaceutical Sciences*, vol. 29, no. 3-4, pp. 269–277, 2006.

[9] C. Hilgendorf, G. Ahlin, A. Seithel, P. Artursson, A.-L. Ungell, and J. Karlsson, "Expression of thirty-six drug transporter genes in human intestine, liver, kidney, and organotypic cell lines," *Drug Metabolism and Disposition*, vol. 35, no. 8, pp. 1333–1340, 2007.

[10] T. G. H. A. Tucker, A. M. Milne, S. Fournel-Gigleux, K. S. Fenner, and M. W. H. Coughtrie, "Absolute immunoquantification of the expression of ABC transporters P-glycoprotein, breast cancer resistance protein and multidrug resistance-associated protein 2 in human liver and duodenum," *Biochemical Pharmacology*, vol. 83, no. 2, pp. 279–285, 2012.

[11] S. Dauchy, F. Dutheil, R. J. Weaver et al., "ABC transporters, cytochromes P450 and their main transcription factors: expression at the human blood-brain barrier," *Journal of Neurochemistry*, vol. 107, no. 6, pp. 1518–1528, 2008.

[12] R. Shawahna, Y. Uchida, X. Declèves et al., "Transcriptomic and quantitative proteomic analysis of transporters and drug metabolizing enzymes in freshly isolated human brain microvessels," *Molecular Pharmaceutics*, vol. 8, no. 4, pp. 1332–1341, 2011.

[13] Y. Uchida, S. Ohtsuki, Y. Katsukura et al., "Quantitative targeted absolute proteomics of human blood-brain barrier transporters and receptors," *Journal of Neurochemistry*, vol. 117, no. 2, pp. 333–345, 2011.

[14] J. F. Deeken and W. Löscher, "The blood-brain barrier and cancer: transporters, treatment, and trojan horses," *Clinical Cancer Research*, vol. 13, no. 6, pp. 1663–1674, 2007.

[15] C. Lhommé, F. Joly, J. L. Walker et al., "Phase III study of valspodar (PSC 833) combined with paclitaxel and carboplatin compared with paclitaxel and carboplatin alone in patients with stage IV or suboptimally debulked stage III epithelial ovarian cancer or primary peritoneal cancer," *Journal of Clinical Oncology*, vol. 26, no. 16, pp. 2674–2682, 2008.

[16] N. Akhtar, A. Ahad, R. K. Khar et al., "The emerging role of P-glycoprotein inhibitors in drug delivery: a patent review," *Expert Opinion on Therapeutic Patents*, vol. 21, no. 4, pp. 561–576, 2011.

[17] A. M. S. Hartz and B. Bauer, "Regulation of ABC transporters at the blood-brain barrier: new targets for CNS therapy," *Molecular Interventions*, vol. 10, no. 5, pp. 293–304, 2010.

[18] L. Milane, S. Ganesh, S. Shah, Z.-F. Duan, and M. Amiji, "Multi-modal strategies for overcoming tumor drug resistance: hypoxia, the Warburg effect, stem cells, and multifunctional nanotechnology," *Journal of Controlled Release*, vol. 155, no. 2, pp. 237–247, 2011.

[19] D. J. Begley, "ABC transporters and the blood-brain barrier," *Current Pharmaceutical Design*, vol. 10, no. 12, pp. 1295–1312, 2004.

[20] A. Ponte-Sucre, M. Padrón-Nieves, and E. Díaz, "ABC transporter blocker and reversal of drug resistance in microorganisms," in *ABC Transporters in Microorganisms. Research, Innovation and Value as Targets against Drug Resistance*, A. Ponte-Sucre, Ed., Caister Academic Press, Norfolk, UK, 2009.

[21] E. Hazai, I. Hazai, I. Ragueneau-Majlessi, S. P. Chung, Z. Bikadi, and Q. Mao, "Predicting substrates of the human breast cancer resistance protein using a support vector machine method.," *BMC Bioinformatics*, vol. 14, p. 130, 2013.

[22] L. Zhong, C.-Y. Ma, H. Zhang et al., "A prediction model of substrates and non-substrates of breast cancer resistance protein (BCRP) developed by GA-CG-SVM method," *Computers in Biology and Medicine*, vol. 41, no. 11, pp. 1006–1013, 2011.

[23] G. F. Ecker, "QSAR studies on ABC transporter—How to deal with polyspecificity," in *Transporters as Drug Carriers*, G. F. Ecker and P. Chiba, Eds., Wiley-VCH, Weinheim, Germany, 2009.

[24] M. A. Demel, O. Krämer, P. Ettmayer, E. E. J. Haaksma, and G. F. Ecker, "Predicting ligand interactions with ABC transporters in ADME," *Chemistry and Biodiversity*, vol. 6, no. 11, pp. 1960–1969, 2009.

[25] J. E. Penzotti, M. L. Lamb, E. Evensen, and P. D. J. Grootenhuis, "A computational ensemble pharmacophore model for identifying substrates of P-glycoprotein," *Journal of Medicinal Chemistry*, vol. 45, no. 9, pp. 1737–1740, 2002.

[26] W.-X. Li, L. Li, J. Eksterowicz, X. B. Ling, and M. Cardozo, "Significance analysis and multiple pharmacophore models for differentiating P-glycoprotein substrates," *Journal of Chemical Information and Modeling*, vol. 47, no. 6, pp. 2429–2438, 2007.

[27] V. Svetnik, T. Wang, C. Tong, A. Liaw, R. P. Sheridan, and Q. Song, "Boosting: an ensemble learning tool for compound classification and QSAR modeling," *Journal of Chemical Information and Modeling*, vol. 45, no. 3, pp. 786–799, 2005.

[28] D.-S. Cao, J.-H. Huang, J. Yan et al., "Kernel k-nearest neighbor algorithm as a flexible SAR modeling tool," *Chemometrics and Intelligent Laboratory Systems*, vol. 114, pp. 19–23, 2012.

[29] N. Giri, S. Agarwal, N. Shaik, G. Pan, Y. Chen, and W. F. Elmquist, "Substrate-dependent breast cancer resistance protein (Bcrp1/Abcg2)-mediated interactions: consideration of multiple binding sites in in vitro assay design," *Drug Metabolism and Disposition*, vol. 37, no. 3, pp. 560–570, 2009.

[30] K. Takenaka, J. A. Morgan, G. L. Scheffer et al., "Substrate overlap between Mrp4 and Abcg2/Bcrp affects purine analogue drug cytotoxicity and tissue distribution," *Cancer Research*, vol. 67, no. 14, pp. 6965–6972, 2007.

[31] E. Hazai and Z. Bikádi, "Homology modeling of breast cancer resistance protein (ABCG2)," *Journal of Structural Biology*, vol. 162, no. 1, pp. 63–74, 2008.

[32] J. D. Allen, S. C. Jackson, and A. H. Schinkel, "A mutation hot spot in the Bcrp1 (Abcg2) multidrug transporter in mouse cell lines selected for doxorubicin resistance," *Cancer Research*, vol. 62, no. 8, pp. 2294–2299, 2002.

[33] C. Özvegy-Laczka, G. Köblös, B. Sarkadi, and A. Váradi, "Single amino acid (482) variants of the ABCG2 multidrug transporter: major differences in transport capacity and substrate recognition," *Biochimica et Biophysica Acta*, vol. 1668, no. 1, pp. 53–63, 2005.

[34] R. W. Robey, Y. Honjo, K. Morisaki et al., "Mutations at amino-acid 482 in the ABCG2 gene affect substrate and antagonist specificity," *British Journal of Cancer*, vol. 89, no. 10, pp. 1971–1978, 2003.

[35] O. Polgar, R. W. Robey, and S. E. Bates, "ABCG2: structure, function and role in drug response," *Expert Opinion on Drug Metabolism and Toxicology*, vol. 4, no. 1, pp. 1–5, 2008.

[36] A. Pozza, J. M. Perez-Victoria, A. Sardo, A. Ahmed-Belkacem, and A. Di Pietro, "Purification of breast cancer resistance protein ABCG2 and role of arginine-482," *Cellular and Molecular Life Sciences*, vol. 63, no. 16, pp. 1912–1922, 2006.

[37] T. Janvilisri, S. Shahi, H. Venter, L. Balakrishnan, and H. W. Van Veen, "Arginine-482 is not essential for transport of antibiotics, primary bile acids and unconjugated sterols by the human breast cancer resistance protein (ABCG2)," *Biochemical Journal*, vol. 385, no. 2, pp. 419–426, 2005.

[38] K. F. K. Ejendal, N. K. Diop, L. C. Schweiger, and C. A. Hrycyna, "The nature of amino acid 482 of human ABCG2 affects substrate transport and ATP hydrolysis but not substrate binding," *Protein Science*, vol. 15, no. 7, pp. 1597–1607, 2006.

[39] L. Eddabra, T. Wenner, H. El Btaouri et al., "Arginine 482 to glycine mutation in ABCG2/BCRP increases etoposide transport and resistance to the drug in HEK-293 cells," *Oncology Reports*, vol. 27, no. 1, pp. 232–237, 2012.

[40] J. Cervenak, H. Andrikovics, C. Özvegy-Laczka et al., "The role of the human ABCG2 multidrug transporter and its variants in cancer therapy and toxicology," *Cancer Letters*, vol. 234, no. 1, pp. 62–72, 2006.

[41] R. Hariharan, A. Janakiraman, R. Nilakantan et al., "MultiMCS: a fast algorithm for the maximum common substructure problem on multiple molecules," *Journal of Chemical Information and Modeling*, vol. 51, no. 4, pp. 788–806, 2011.

[42] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, "Optimization clustering techniques," in *Cluster Analysis*, D. J. Balding, N. A. C. Cressie, G. M. Fitzmaurice et al., Eds., John Wiley & Sons, Chichester, UK, 5th edition, 2011.

[43] A. Varnek and I. Baskin, "Machine learning methods for property prediction in chemoinformatics: quo vadis?" *Journal of Chemical Information and Modeling*, vol. 52, pp. 1413–1437, 2012.

[44] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832–844, 1998.

[45] A. Talevi, C. L. Bellera, M. Di Ianni, P. R. Duchowicz, L. E. Bruno-Blanch, and E. A. Castro, "An integrated drug development approach applying topological descriptors," *Current Computer-Aided Drug Design*, vol. 8, no. 3, pp. 172–181, 2012.

[46] N. Triballeau, F. Acher, I. Brabet, J.-P. Pin, and H.-O. Bertrand, "Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4," *Journal of Medicinal Chemistry*, vol. 48, no. 7, pp. 2534–2547, 2005.

[47] R. Hubbard and M. J. Bayarri, "Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing," *American Statistician*, vol. 57, no. 3, pp. 171–178, 2003.

[48] J.-F. Truchon and C. I. Bayly, "Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem," *Journal of Chemical Information and Modeling*, vol. 47, no. 2, pp. 488–508, 2007.

[49] L. E. Dodd and M. S. Pepe, "Partial AUC estimation and regression," *Biometrics*, vol. 59, no. 3, pp. 614–623, 2003.

[50] J. Kirchmair, P. Markt, S. Distinto, G. Wolber, and T. Langer, "Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection—what can we learn from earlier mistakes?" *Journal of Computer-Aided Molecular Design*, vol. 22, no. 3-4, pp. 213–228, 2008.

[51] W. Zhao, K. E. Hevener, S. W. White, R. E. Lee, and J. M. Boyett, "A statistical framework to evaluate virtual screening," *BMC Bioinformatics*, vol. 10, article 225, 2009.

[52] D. Bamber, "The area above the ordinal dominance graph and the area below the receiver operating characteristic graph," *Journal of Mathematical Psychology*, vol. 12, no. 4, pp. 387–415, 1975.

[53] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[54] L. L. Pesce, J. Papaioannu, and C. E. Metz, "ROC-kit software 2009," http://metz-roc.uchicago.edu/MetzROC/software.

[55] X. Robin, N. Turck, A. Hainard et al., "pROC: an open-source package for R and S+ to analyze and compare ROC curves," *BMC Bioinformatics*, vol. 12, article 77, 2011.

[56] M. S. Pepe, *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford University Press, Oxford, UK, 2004.

[57] L. L. Pesce, C. E. Metz, and K. S. Berbaum, "On the convexity of ROC curves estimated from radiological test results," *Academic Radiology*, vol. 17, no. 8, pp. 960–968.e4, 2010.

[58] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," *Biometrics*, vol. 44, no. 3, pp. 837–845, 1988.

[59] J. Carpenter and J. Bithell, "Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians," *Statistics in Medicine*, vol. 19, no. 9, pp. 1141–1164, 2000.

[60] V. Kairys, M. X. Fernandes, and M. K. Gilson, "Screening drug-like compounds by docking to homology models: a systematic study," *Journal of Chemical Information and Modeling*, vol. 46, no. 1, pp. 365–379, 2006.

[61] S. K. Kearsley, S. Sallamack, E. M. Fluder, J. D. Andose, R. T. Mosley, and R. P. Sheridan, "Chemical similarity using physiochemical property descriptors," *Journal of Chemical Information and Computer Sciences*, vol. 36, no. 1, pp. 118–127, 1996.

[62] R. P. Sheridan, S. B. Singh, E. M. Fluder, and S. K. Kearsley, "Protocols for bridging the peptide to nonpeptide gap in topological similarity searches," *Journal of Chemical Information and Computer Sciences*, vol. 41, no. 3–6, pp. 1395–1406, 2001.

[63] H. Yabuuchi, S. Niijima, H. Takematsu et al., "Analysis of multiple compound-protein interactions reveals novel bioactive molecules," *Molecular Systems Biology*, vol. 7, article 472, 2011.

[64] I. V. Tetko, "Neural network studies. 4. Introduction to associative neural networks," *Journal of Chemical Information and Computer Sciences*, vol. 42, no. 3, pp. 717–728, 2002.

[65] I. V. Tetko, V. Y. Tanchuk, N. P. Chentsova et al., "HIV-1 reverse transcriptase inhibitor design using artificial neural networks," *Journal of Medicinal Chemistry*, vol. 37, no. 16, pp. 2520–2526, 1994.

[66] N. V. Artemenko, I. I. Baskin, V. A. Palyulin, and N. S. Zefirov, "Artificial neural network and fragmental approach in prediction of physicochemical properties of organic compounds," *Russian Chemical Bulletin*, vol. 52, no. 1, pp. 20–29, 2003.

[67] N. I. Zhokhova, I. I. Baskin, V. A. Palyulin, A. N. Zefirov, and N. S. Zefirov, "Fragmental descriptors with labeled atoms and their application in QSAR/QSPR studies," *Doklady Chemistry*, vol. 417, no. 2, pp. 282–284, 2007.

[68] H. Zhu, A. Tropsha, D. Fourches et al., "Combinatorial QSAR modeling of chemical toxicants tested against Tetrahymena pyriformis," *Journal of Chemical Information and Modeling*, vol. 48, no. 4, pp. 766–784, 2008.

[69] A. Varnek, D. Fourches, D. Horvath et al., "ISIDA: platform for virtual screening based on fragment and pharmacophoric descriptors," *Current Computer-Aided Drug Design*, vol. 4, no. 3, pp. 191–198, 2008.